1 Exploratory data analysis

1. Describe your data
2. Domain context of the data
3. Summaries of variables
4. Boxplots and histograms of variables
5. Potential outliers and possible domain reasons if information is available
6. Pair plots when the number of predictors is not too large. If too large for pair plot, a few important ones.
7. Correlations between predictors
   (a) Correlation matrix image
   (b) Table of correlation and CI for each pair of variables
   (c) Interpretation
8. Correlations between response and predictors
   (a) Table of correlation and CI between response and each predictor
   (b) Interpretation
9. Questions you want answered through linear regression analysis

2 Linear regression analysis

1. Obtain a fit for the full model (i.e., all predictors included)
2. Fit summary and its interpretation: e.g., confidence intervals, p-values, etc.
3. Plot of data and fit for simple linear regression together with confidence band

3 Diagnostics

1. Inspect and interpret plots produced by plot(lm(...))
2. Look for signs for problems:
   (a) Outliers
      i. Non-influential: don't worry about it.
      ii. Influential: remove (or use robust regression, but not necessary for this miniproject.)
      iii. See if the data domain offers any explanation for these
   (b) Influential points
      i. Fit regression model with and without the point and report both analyses.
      ii. See if the data domain offers any explanation for these
   (c) Nonconstant variance
      i. Use transformation or nonparametric methods.
      ii. Note: doesn't affect the fit too much; mainly an issue for confidence intervals.
   (d) Nonlinearity: Use transformation or nonparametric methods.
   (e) Nonnormality
      i. Large samples: not a problem.
      ii. Small samples: use transformations (e.g., log. More generally Box-Cox but this need not be done for this miniproject.)

4 Model selection

Apply any method for model/feature/variable selection and summarize the model selected thereby.